心理科学进展 2021, Vol. 29, No. 12, 2172–2183
Advances in Psychological Science

https://doi.org/10.3724/SP.J.1042.2021.02172

自动驾驶车中的人机信任*

高在峰[1] 李文敏[1] 梁佳文[1] 潘晗希[1] 许 为[2] 沈模卫[1]
([1]浙江大学心理与行为科学系, [2]浙江大学心理科学中心, 杭州 310007)

摘 要 自动驾驶是当前智能汽车发展的重要方向。在实现完全自动化驾驶前, 驾驶员和自动驾驶系统共享车辆控制权, 协同完成驾驶任务。在该人−机共驾阶段, 人对自动驾驶系统的信任是影响自动驾驶中人机协同效率与驾驶安全的关键要素; 驾驶员对自动驾驶车辆保持适当的信任水平对驾驶安全至关重要。本研究结合信任的发展阶段与影响因素提出了动态信任框架。该框架将信任发展分为倾向性信任、初始信任、实时信任和事后信任四个发展阶段, 并结合操作者特征(人)、系统特征(自动驾驶车系统)、情境特征(环境)三个关键因素分析不同阶段的核心影响因素以及彼此间的内在关联。根据该框架, 信任校准可从监测矫正、驾驶员训练、优化 HMI 设计三类途径展开。未来研究应更多关注驾驶员和人机系统设计特征对信任的影响, 考察信任的实时测量和功能特异性, 探讨驾驶员和系统的相互信任机制, 以及提升信任研究的外部效度。

关键词 信任, 自动驾驶, 动态信任框架, 信任校准, HMI 设计
分类号 B849: U491

1 引言

自动驾驶(Automated Driving, AD)是新一轮科技革命中的典型代表, 在改变人类出行方式、解决道路拥堵和安全问题等方面将起到重要作用, 是诸多国家的战略发展方向(DOT, 2018; ERTRAC, 2019; SAE China, 2020)。目前, 美国汽车工程师学会(Society of Automotive Engineers, SAE)对自动驾驶车辆的自动化程度等级划分被广为接受, 它将汽车的自动化程度从无自动化(L0)到完全自动化(L5)划分为 6 个等级(SAE, 2018)。从无自动化的手动驾驶到完全自动化的无人驾驶, 操作者的角色逐步从驾驶者转变为监管者, 驾驶员与车辆的关系亦随之从控制与被控制的关系转变为双向合作关系。在实现 L5 无人驾驶前, 驾驶员和自动驾驶系统共享车辆控制权, 协同完成驾驶任务, 该状态被称为人−机共驾(Human-machine Cooperative Driving)。信任(Trust)在自动驾驶环境下的人−机共驾中(人机信任, trust in automation)扮演重要角色(Hancock et al., 2019; Rahwan et al., 2019), 是影响自动驾驶中人机协同效率与驾驶安全的关键要素(Hancock et al., 2019; Rahwan et al., 2019)。若驾驶员不信任自动系统, 可能会忽视系统提供的辅助功能, 无法有效降低疲劳驾驶、分心等交通风险; 若驾驶员过度信任系统, 则会放弃对车辆的监控而导致巨大安全隐患(如 Kelleher, 2018; Noah et al., 2017; Noah & Walker, 2017; NTSB, 2018; Wintersberger et al., 2018)。

因此, 自动驾驶中的信任问题成为近年来的研究热点(如 Choi & Ji, 2015; Ekman et al., 2018; Hergeth et al., 2017; Koo et al., 2015; Mishler, 2019; Molnar et al., 2018; Payre et al., 2016; Verberne et al., 2012), 积累了一批重要成果。研究者亟需一个模型或框架来整合已有研究发现, 阐述自动驾驶环境下的信任动态变化过程及其相关影响机制。然而, 已有信任模型或关注一般自动化系统的静态信任结构及其相关影响因素(如 Adams et al., 2003; Chien, Lewis, et al., 2014; Hancock et al., 2011; Hoff & Bashir, 2015; Lee & See, 2004; Schaefer et al., 2014), 或仅关注自动驾驶的实时信任而忽略了信任的其他发展阶段。为此, 本文

收稿日期: 2020-12-27
* 科技创新 2030 子课题(2018AAA0101605); 科技部重点研发计划(2019YFB1600504)。
通信作者: 高在峰, E-mail: zaifengg@zju.edu.cn

将系统梳理自动驾驶车领域的信任研究，阐述自动驾驶信任的内涵并提出信任的动态发展框架，最后基于该框架讨论信任的校准与未来研究方向。

1 人机信任的内涵

Lee 和 See (2004)提出的人机信任定义被研究者广为接受(如 French et al., 2018; Hoff & Bashir, 2015; Khastgir et al., 2017)。他们认为态度、意愿与行为间存在内在联系：操作者对系统的态度影响其使用系统的意愿和依赖行为，但是依赖系统和有使用系统意愿并不代表信任系统。故 Lee 和 See 从态度角度定义信任，即信任是个体(如驾驶员)在不确定或易受伤害的情境下认为代理(agent, 如自动驾驶系统)能帮助其实现某个目标(如驾驶任务)的态度。

1.1 信任测量

有关自动驾驶的信任测量围绕信任的内涵来开展，主要从驾驶员对自动驾驶系统的依赖行为、生理基础、驾驶员的主观态度等方面着手。具体来讲，第一，测量驾驶员对自动驾驶系统的依赖行为，如驾驶过程中手动驾驶与自动驾驶占总体时间比率(如 de Vries et al., 2003; Molnar et al., 2018)、驾驶员接管后归还自动驾驶控制权的延迟时间(如 Hergeth et al., 2015; Molnar, 2017)、驾驶员对驾驶相关区域的监控频率或监控时间占总时间的比率(Hergeth et al., 2016; Li et al., 2020)等。第二，测量驾驶员在驾驶过程中的心率、皮肤电等生理指标。当自动驾驶系统发出接管请求时，若驾驶员信任系统，其情绪状态将较为稳定，心率和心率变异性将更为平缓(如 Petersen et al., 2017; Waytz et al., 2014)，皮肤电活动水平较低(如 Morris et al., 2017)；反之，若不信任系统，驾驶员将会紧张焦虑进而影响生理指标。第三，测量驾驶员对自动驾驶系统的主观信任程度(如 Brown & Galster, 2004; Chien, Lewis, et al., 2014; Chien, Semnani-Azad, et al., 2014; Jian et al., 2000; Soh et al., 2009; Körber et al., 2018)。如 Jian 等人(2000)的人机信任问卷是目前使用最广泛的问卷(如 Beggiato & Krems, 2013; Gold et al., 2015; Niu et al., 2018; Verberne et al., 2012)。

1.2 信任校准

衡量系统能力与驾驶员实际信任水平的关系，是人机信任领域的核心问题之一。研究者一般通过以系统能力(Capability)为横轴、信任水平(Trust)为纵轴的二维坐标系(图 1)来描述二者间关系(de Visser et al., 2014; Lee & See, 2004)。系统能力反映操作者在特定情境下基于该系统能力应具备的客观可信任水平(Trustworthiness)，而信任水平反映在实际人机交互中操作者对系统的*主观实际*信任水平。我们可通过衡量主观实际信任与客观可信任水平间的相对关系，来评估当前的信任状态是否适当，进而校准信任(Lee & See, 2004; Walker & Stanton, 2017)。

信任是否适当，一般可通过三个方面来评估(Lee & See, 2004)。第一，主观实际信任水平和客观可信任水平间的匹配(Calibration)。根据二者的相对关系，信任存在适当信任(Appropriate trust)、信任不足(Under-trust)和过度信任(Over- trust)三种状态。适当信任也称校准的信任(Calibrated trust)，指驾驶员的主观实际信任水平与客观可信任水平一致，如图 1 对角线所示(de Visser et al., 2014)。信任不足指驾驶员的主观实际信任水平低于客观可信任水平(图 1 右下方区域)，通常是因为驾驶员低估自动驾驶系统的能力，此时驾驶员往往会忽视系统提供的有效建议，导致驾驶员不使用自动系统功能(Disuse)。过度信任指驾驶员的主观实际信任水平高于客观可信任水平(图 1 左上方区域)，通常是因为驾驶员高估自动驾驶系统的能力，此时

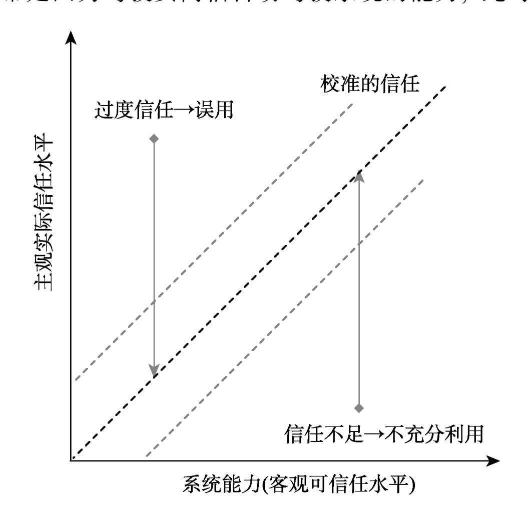


图 1 系统能力与主观实际信任水平间的关系(改编于 de Visser et al., 2014; Lee & See, 2004)。图中灰色虚线区域表示在实际应用中，信任校准存在一个信任水平不合适但可恢复或安全的区域(具体范围有待进一步探讨)。

驾驶员往往不会及时监控当前的车况和路况，导致驾驶员滥用(Misuse)自动系统功能。第二，主观实际信任对客观可信任水平的分辨率(Resolution)。高分辨率指当系统客观可信任水平变化较大时，主观实际信任水平随之发生较大变化；低分辨率指客观可信任水平变化较大时，主观实际信任水平未发生变化或变化幅度较小。第三，主观实际信任的特异性(Specificity)，又可分为时间特异性与功能特异性两类。时间特异性指驾驶员主观实际信任随客观可信任水平实时变化的程度。高时间特异性代表高时间灵敏度：系统出现失误时主观实际信任随之下降，反之则存在时间滞后性。功能特异性指驾驶员对自动驾驶系统不同子系统、功能模块、驾驶模式有不同的信任程度。高功能特异性反映驾驶员对不同子系统等有差异化的信任水平，低功能特异性则相反。目前有关自动驾驶的人机信任研究主要围绕第一方面展开，即人–机共驾中如何规避信任不足与过度信任，帮助驾驶员达到或维持适当信任水平。

2 动态信任框架

我们以信任的发展过程为主线，在系统梳理信任不同发展阶段的影响因素及其内在逻辑关系的基础上，提出了基于信任发展全过程的自动驾驶动态信任框架(见图 2)，以阐述自动驾驶中信任的发展过程(即动态的含义所在)及相关影响机制(Chen et al., 2021)。该部分将详细介绍该框架。需指出，本框架适用于不同等级(非 L0)的自动驾驶系统，相关要素的影响作用在不同等级的系统中可能会有所变化。

2.1 框架要素

从信任的发展过程来看，Merritt 和 Ilgen (2008)认为操作者对自动化系统的信任是一个处于倾向性信任(Dispositional trust)与历史性信任(History-based trust)间的连续体，前者属于人对自动系统信任的固有倾向，后者属于人通过与自动系统交互后而形成的信任状态；操作者的信任水平时刻处于该连续体的某一点上。随着系统的持续使用，操作者对系统的信任逐渐从以倾向性信任成分为主转变为以历史性信任成分为主，并历经初始信任(Initial trust)、实时信任(Ongoing trust)和事后信任(Post-task trust)三种历史信任状态(French et al., 2018; Merritt & Ilgen, 2008)。因此，本框架将信任发展分为 4 个发展阶段：倾向性信任、初始信任、实时信任和事后信任。初始信任在倾向性信任基础上发展而来，指操作者对即将使用的自动系统已具备一定认知、但尚未使用前，对其所持有的信任状态；实时信任指操作者在人机交互过程中对系统的信任状态；而事后信任则指操作者在结束交互后对系统的信任状态，属于事后对系统信任的总体评估。事后信任与实时信任的影响因素高度重合，因此本模型主要关注倾向性信任、初始信任和实时信任的影响因素。

在信任的影响因素方面，本文认为主要包括操作者特征(人)、系统特征(自动驾驶车系统)、情境特征(环境)三个方面。其中，操作者特征可划分为固有特质与先验经验两种因素。固有特质指个体生理固有或长期形成的相对稳定特征，如性

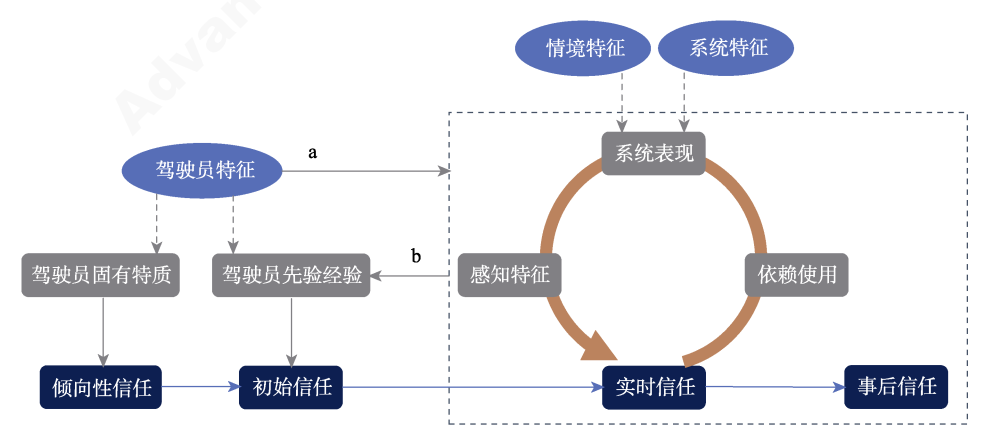


图 2 基于信任发展过程的自动驾驶动态信任框架。a 线表示驾驶员特征影响除系统表现外的其他所有四个因素，b 线表示框中所有因素均可转化为驾驶员的先验经验。

别、人格、年龄、文化背景等因素，与系统和情境无关；先验经验指通过学习而获得的系统和情境特征信息，在很大程度上反映了驾驶员对车的掌握能力。在驾驶员使用自动驾驶系统的过程中，系统特征与情境特征通过系统表现客观地反映出来，而客观的系统表现特征经驾驶员认知系统加工后转为主观感知特征(亦包含了个体对系统表现的潜在风险感知)，后者是影响信任的直接因素。

2.2 不同类型信任的影响机制

2.2.1 倾向性信任与初始信任

倾向性信任(Dispositional trust)是信任发展的最初阶段，反映人生来就有的信任倾向(Trust propensity)。倾向性信任主要受驾驶员固有特质(如年龄、人格等)影响。研究发现，驾驶员年龄会影响倾向性信任(Donmez et al., 2006; Gold et al., 2015; Molnar, 2017; Molnar et al., 2018)，年老驾驶员更倾向于信任和使用自动驾驶系统。在使用自动驾驶系统时，驾驶员对系统的控制程度降低，因此驾驶员对系统的控制偏好程度会影响信任：驾驶控制偏好程度越低，驾驶员越倾向于信任自动驾驶系统(Molnar et al., 2018)。驾驶员的人格特质也会影响信任倾向(Chien et al., 2016; Szalma & Taylor, 2011)，如操作者宜人性越高，越倾向于信任自动化系统，这可能与宜人性本身包含信任、依从等特质有关。尚未有研究发现性别对倾向性信任有显著影响(Molnar et al., 2018)。

初始信任(Initial trust)由倾向性信任发展而来。它除受驾驶员已有的倾向性信任影响外，还受驾驶员有关自动系统的先验经验影响。先验信息能促进驾驶员对系统的了解，影响驾驶员有关系统的心理模型，从而影响信任。在初次接触系统时，先验信息主要来源于他人描述(尤其是因广告而产生的品牌声誉)和相似系统的使用经验。驾驶员对高品牌声誉的自动驾驶系统信任水平更高(Carlson et al., 2014; Celmer et al., 2018)；有关系统是否可能出错的提示信息会影响驾驶员的初始信任水平(Beggiato, & Krems, 2013)；自动化程度越高，驾驶员对系统的初始信任水平可能也越高。在多次接触系统后，驾驶员与自动系统的交互经验会转化为驾驶员的先验知识，影响下一次驾驶的初始信任(如图 2 中的 b 线所示)。研究发现，接管经验可让驾驶员体验到系统的不足，利于驾驶员更好的了解系统、进而提高信任水平(Gold et al., 2015; Hergeth et al., 2017; Molnar et al., 2018)。需要指出，接管经验对信任的影响可能并非是单一的，如接管体验在短期内暴露了系统的不足导致信任降低，但从长期看会增进驾驶员对自动化系统的了解，对系统的能力和限制将会有更为精确的心理模型，进而促进合适信任水平的达成(Hergeth et al., 2015; Payre et al., 2016)。

2.2.2 实时信任

实时信任(Ongoing trust)由初始信任发展而来。驾驶员的实时信任水平会直接影响驾驶员是否使用自动化系统以及在多大程度上依赖它。在使用自动化系统的过程中，自动化系统的系统特征和驾驶相关的情境特征是客观特征，是系统表现的直接影响因素。客观特征决定驾驶员对自动化系统应持有的客观可信任水平；但客观特征不会直接影响实时信任，需驾驶员通过认知加工转化为主观感知特征后方会影响实时信任。故主观感知特征是信任的直接影响因素，如图 2 所示。

实时信任主要受系统特征和情境特征影响。系统特征包括系统目的、系统过程和系统能力。系统目的指设计者或系统功能的设计目的，如跟车巡航、辅助变道等。用户通常会认为系统提供的功能是可靠的，并对其产生信任(Hoff & Bashir, 2015)。系统过程指自动化系统完成驾驶任务的方式。研究发现，自动驾驶系统在预警中提供驾驶相关信息有助于增强信任(Koo et al., 2015; Verberne et al., 2012)；提供驾驶指导比只提供车辆信息的驾驶辅助系统更值得信任(Cramer et al., 2008)；事前反馈比事后反馈的系统有更高的信任水平(Du et al., 2019)。系统能力是影响信任的重要因素，如系统可靠性(Hergeth et al., 2017; Jonsson et al., 2008; Petersen et al., 2018)、系统错误(Kraus et al., 2020)、系统的自动化水平等。系统错误越严重、可靠性越低，信任受损越严重；系统失效(漏报)会严重损害对系统的信任(Mishler, 2019)。情境特征包括任务难度、路况和天气等。已有研究发现，车流密度影响驾驶员对系统的信任水平(施彦玮, 2019)。

主观感知特征是影响实时信任的核心要素，因此恰当的实时信任取决于驾驶员对系统、情境特征的准确感知(即驾驶员的情景意识水平；Endsley, 1995, 2016)。如图 2 中的含箭头圆环所示，实时信任影响驾驶员对系统的依赖使用行为，在

自动驾驶过程中驾驶员通过系统表现可动态感知到系统特征与情境特征，并据此动态调整其实时信任水平。当系统表现持续保持良好时，驾驶员对系统的信任水平会逐渐提高，而当系统出错后驾驶员的信任水平将降低(Kraus et al., 2020; Mishler, 2019)。系统亦可通过人机界面(Human-Machine Interface, HMI)提供有关系统特征和情境特征的信息以提升驾驶员的准确感知能力。在系统特征方面，研究发现通过视觉方式呈现系统的可靠性或不确定性信息(如以柱状图形式告知当前情景下自动驾驶系统的应对能力)有助于驾驶员形成适当的信任水平，避免过度信任(Helldin et al., 2013; Kunze et al., 2019)。在情境特征方面，环境重构视图可以让驾驶员更准确地感知环境风险，使其信任水平随实际交通风险水平而动态调整(施彦玮, 2019)。

近期研究开始关注HMI中存在的社会线索对信任的影响。自动化系统携带的相关社会线索可从自动驾驶系统的特定外观、驾驶行为和决策模式等方面被驾驶员感知到，从而影响驾驶员对系统的信任。这些社会线索包含拟人化、自动系统−驾驶员相似性、驾驶风格等。拟人化指赋予自动驾驶系统以拟人化的特征(如语音、外观、性别)。拟人化特征通过增强驾驶员对自动化系统的理解(Niu et al., 2018)、情感联系(Epley et al., 2007; Häuslschmid et al., 2017)或社会临场感(Social presence, Lee et al., 2015)来提高驾驶员的信任水平(如 Forster et al., 2017; Häuslschmid et al., 2017; Waytz et al., 2014; Zihsler et al., 2016)。自动化系统−驾驶员相似性包括外观相似性、行为相似性和认知相似性三方面。自动化系统与驾驶员间的认知相似性(如有共同的驾驶目标)能提高驾驶员对系统的信任水平(Verberne et al., 2012; Verberne et al., 2015)，这可能是由于相似相吸(Verberne et al., 2015)。自动系统驾驶风格类似人的驾驶风格，可体现在变道、加速、刹车、跟车距离、横向安全距离等行为模式上。研究发现，驾驶员对保守驾驶风格的自动驾驶车辆信任水平高于对激进驾驶风格车辆的信任水平(Ekman et al., 2019)，这可能是因为驾驶员从前者感知到的风险较低。若同时考虑驾驶员和车辆的驾驶风格，Hartwich 等人(2018)发现驾驶风格与驾驶员相似的自动驾驶系统更值得信任。

需指出，在客观特征转化为主观感知特征的过程中，驾驶员特征会起调节作用从而影响信任(如图 2 中的 a 线所示)。Li 等人(2020)发现，固有人格特质影响驾驶员的实时信任水平。如低开放性特质较高开放性特质的驾驶员在驾驶中具有更高的实时信任水平，这可能是因为人格特质影响了个体的监控行为，导致不同性格驾驶员对系统的特征感知存在差异。其他操作者特征(如情绪状态)可能也会影响操作者对客观特征的感知与转化，但尚未见相关研究。驾驶员先验经验会影响其对系统运行的心理模型或预期。当实际系统表现与驾驶员的预期不一致时，信任水平可能会下降(Kraus et al., 2020; Zhang et al., 2018)。此外，驾驶员对系统的初始信任水平越高，其对系统的预期也越高，在遇到系统失效时信任降低也将更为严重(Merritt & Ilgen, 2008)；给驾驶员提供有关系统错误的不同类型信息提示时(如准确提示、没有提示、错误提示)，准确提示系统可能出现失误时驾驶员的信任水平最高(Beggiato, & Krems, 2013)。总体来讲，驾驶员特征对实时信任的影响研究关注较少，需要加强。

3 信任校准

研究信任的最终目的在于信任校准，使驾驶员对自动化系统保持恰当信任水平。信任校准实质上是要确保实时信任处于合理的水平(即图 1 的对角线)。基于上文的自动驾驶动态信任框架，信任校准可从监测矫正、驾驶员训练、优化 HMI 设计三个方面入手。

3.1 监测矫正

校准不当信任的最直接方法是监测驾驶员的实时信任水平：当驾驶员对系统信任不足或过度信任时给予适当干预。当驾驶员处于过度信任时，系统可向驾驶员提供告警反馈、系统可靠性信息等以提示当前系统风险和环境风险，校准驾驶员的不适当认知，完成驾驶员对系统信任水平的调整(如 Helldin et al., 2013; Kunze et al., 2019)。当驾驶员处于信任不足时，可通过向驾驶员提供系统可靠性信息等来提高其信任(如 Helldin et al., 2013; Kunze et al., 2019)，亦可通过视觉、听觉等形式的信息反馈来实现。人际关系中的道歉、否认、解释、承诺等方式可用于修复驾驶员与自动驾驶系统间的信任(trust recovery, de Visser et al.,

2018; Khastgir et al., 2017; Kohn et al., 2018)。

当前的监测矫正实施存在两个难点。(1)如何动态测量实时信任。实验室研究中的实时信任监测可通过测量驾驶中的行为(如眼动)与生理指标(如心率、皮肤电)来实现。然而, 由于环境光线、肢体动作、情绪状态等因素限制, 在实车上精准测量与信任直接相关的眼动与皮肤电等指标仍存在一定难度; 未来需加强实时测量方法和指标的研究。(2)如何识别不适当信任。信任是否适当的核心在于系统能力和当前实际信任水平间是否匹配。然而, 二者并不在同一个测量维度上。即使给定系统能力和操作者的实时信任水平, 判断此时的信任状态是否合适也有一定难度。可能的解决途径是构建系统能力到客观可信任水平间的映射。例如, 研究者可尝试构建出不同自动驾驶等级与场景下的客观可信任水平。如 L3 级车辆在高速公路场景下执行自动行驶时, 驾驶员对自动系统须达到最低监控频率或最低监控时间, 否则进行干预。总体而言, 监测矫正的实施需在实车的信任指标选择与不恰当信任识别两方面做进一步探索。

3.2　驾驶员训练

根据动态信任框架, 驾驶员的先验经验决定了其对自动驾驶系统的心理模型或预期, 进而影响实时信任的动态变化过程(如图 2 中的 a 线)。因此, 校准不当信任可通过训练来积累驾驶员的先验经验, 促使驾驶员对系统形成正确认知。驾驶员通过训练可了解系统的功能及其不足, 习得利用系统获取环境信息的能力, 提前形成有关自动系统的正确心理模型(Ekman et al., 2018)、降低系统首次失败的影响(first failure, Manzey et al., 2012), 从而提高维持合适信任水平的能力(Gold et al., 2015; Hergeth et al., 2017; Molnar et al., 2018)。此外, 深度训练(经历超车、被超车和接管等复杂场景)可消除过度信任对接管反应的消极影响(Payre et al., 2016)。因此, 有研究者提出将自动驾驶训练(如控制权切换操作等)纳入驾照考核(Saffarian et al., 2012; Toffetti et al., 2009)。然而, 已有研究涉及的均为短暂驾驶体验, 并非真正意义上的系统性训练; 同时, 通过训练所形成的心理模型可能具有较低的生态效度。今后研究需考虑长期训练对自动驾驶的驾驶员影响(Cohen et al., 1997), 并考察基于真实自动驾驶场景训练的作用机制。

3.3　优化 HMI 设计

根据动态信任框架, 驾驶员可通过系统表现来感知系统特征。因此, 校准不当信任可通过优化 HMI 设计来实现, 向驾驶员提供有关系统特征与情境特征的信息, 以提升驾驶员的准确感知能力。本途径侧重通过向驾驶员提供有关系统和情境的信息来提高客观特征的透明度(Transparency)和可理解性(许为, 2019, 2020; Chen et al., 2018), 帮助驾驶员更好的加工客观特征, 增强驾驶员对系统和情境的感知、理解和预测程度(即情境意识), 以校准信任。

HMI 应提供哪些信息是本途径的关键。de Visser 等人(2014)和 Mirnig 等人(2016)就 HMI 应提供的信息分别提出了信任线索的 HMI 设计框架。这两个设计框架均包含系统特征和驾驶员信息加工两个维度。在系统特征维度方面, de Visser 等人从信任影响来源出发, 提出 HMI 需提供系统目的、系统能力、系统过程、表现形式和系统的设计背景与声誉五个维度的系统信息; Miring 等人从系统功能自动化的层次出发, 认为 HMI 需提供操作、战术和战略三个层次的系统信息。其中, de Visser 等人(2014)侧重于告知驾驶员自动驾驶系统的信息, Miring 等人(2016)侧重于告知驾驶员系统将会如何驾驶, 后者可认为是系统过程维度的细化。需注意的是, de Visser 等人所提的五个维度间存在一定交叉, 可进一步简化为系统目的、系统能力和系统过程三个方面, 这与本文模型框架解释部分(见 2.2.2)所阐述的系统特征要素一致。例如, 为传达系统能力信息, 可向驾驶员呈现传感器识别结果或在不同天气下传感器的可靠性; 为传达系统过程信息, 可向驾驶员呈现超车、变道等过程示意图。在驾驶员信息加工维度方面, 已有的两个架构(de Visser et al., 2014; Mirnig et al., 2016)实质上均围绕如何增强驾驶员对系统的情境意识三层次展开(Chen et al., 2018; Endsley, 1995, 2016), 即感知、理解和预测。

此外, 情境特征信息也可能会影响适当信任的形成, 而上述两模型均忽略了该因素应如何在 HMI 中体现。同时, 信任线索的 HMI 框架还需考虑驾驶员特征(Li et al., 2020; Mirnig et al., 2016), 即针对不同类型或不同状态的驾驶员采用不同的 HMI 设计思路, 实现人机环融合。

4 研究展望

基于上述动态信任框架与相关分析，我们认为在自动驾驶车的人机信任方面，以下六个方面的研究值得进一步关注。

(1)驾驶员特征对信任的影响。已有研究主要集中在驾驶员特征对倾向性信任和初始信任的影响方面，且主要关注了年龄、人格的影响。这些研究对我们全面理解人-机共驾下的主体——驾驶员远远不足。在倾向性信任与初始信任方面，尚未有研究探讨不同的文化背景(如集体主义 vs. 个人主义; Chien et al., 2016)、认知风格(整体型 vs. 分析型; Armstrong et al., 2012; Riding & Cheema, 1991)等特质对信任倾向的影响。在实时信任方面，根据本文的动态信任框架，驾驶员特征会通过影响客观系统特征与情境特征的感知而影响实时信任。然而，目前尚未有研究探讨心理负荷、情绪状态等驾驶员特征对客观信息感知的影响。

(2)人机系统设计对实时信任的影响。目前的信任研究主要关注某种特定的HMI设计是否提高了信任，后续研究需注意以下三方面的探讨。第一，HMI 设计的最终目的不是提高信任，而是帮助驾驶员维持合适的信任水平(即校准信任)。因此，研究不仅应关注 HMI 的积极作用，还需考察 HMI 设计是否会引发过度信任。第二，采用认知建模(如 ACT-R; Anderson et al., 2004; Cao et al., 2013)等方式，从整体上把握人机系统设计因素对驾驶员信任的影响，或建立信任动态预测的量化模型(Gao & Lee, 2006)。第三，人机系统设计不仅仅包括 HMI 设计，还包括车辆驾驶行为设计等系统因素，例如道德困境下的决策方式(Awad et al., 2018)。

(3)实时信任的测量。信任校准(尤其是信任的时间特异性)很大程度上取决于我们能否准确监测驾驶员的实时信任水平。目前，实时信任测量主要通过测量驾驶行为和眼动来实现(Hergeth et al., 2016; Walker et al., 2018)，基于生理指标的测量相对较少。基于脑电技术的模式识别(Fahrenfort et al., 2018; Haxby et al., 2014; Meyers, 2013)等技术日渐成熟，可实现便携、无创在线数据采集，在实现快速有效监测驾驶员的实时信任方面具有较为广阔的空间。

(4)驾驶员对自动驾驶系统主观实际信任的功能特异性。目前，有关信任的自动驾驶研究处于起步阶段，大多将自动驾驶系统作为一个整体进行研究，这对研究者与设计者从总体上把握影响信任的因素具有重要意义。然而如前文所指出，驾驶员在实际操作过程中会对自动驾驶系统的不同子系统、功能模块等产生不同的信任水平。目前尚未有研究就信任的功能特异性进行探讨。该方面的研究对精准校正驾驶员的自动驾驶信任具有重要价值。

(5)驾驶员和自动系统的相互信任。当系统智能化程度逐渐提高，具备认知推理和自适应能力，自动系统将从自动化转为自主化，成为智能体。此时，人机信任将不再是单向的人对自动系统的信任，而是双向的，即人机互信(许为, 2019)。在人机互信框架下，以下两个方面的研究亟需开展。第一，系统对驾驶员的信任。研究者可基于驾驶员当前状态(疲劳、分心等)、系统状态(可靠性等)和情境状况(环境风险等)等数据进行建模，构建系统对驾驶员的适当信任模型，系统在驾驶员处于不可信状态时主动介入以避免事故。第二，团队信任修复。在复杂路况或驾驶者未及时接管时，自动驾驶可能会发生意外，从而损害了驾驶员对自动系统的信任(de Visser et al., 2018; Khastgir et al., 2017; Kohn et al., 2018)。意外发生后系统应通过何种方式来主动修复驾驶员对系统的信任(de Visser et al., 2018; Khastgir et al., 2017)对协同驾驶而言很重要。

(6)提升信任研究的外部效度。目前的信任研究大多在实验室中完成，研究的外部效度相对较低，研究结果不一定能直接应用。因此，有必要在保障安全的情况下，加强实车道路研究。如研究者可通过 Wizard of Oz 实验方法，让已训练过的操作员在车辆的后排座位或远程控制实际道路上的车辆，以达到模拟自动驾驶效果(如 Ekman et al., 2016; Ekman et al., 2019)。

参考文献

施彦玮. (2019). *环境知觉对 L2 自动驾驶人机信任的影响* (硕士学位论文). 浙江大学, 杭州.

许为. (2019). 四论以用户为中心的设计：以人为中心的人工智能. *应用心理学*, *25*(4), 291–305.

许为. (2020). 五论以用户为中心的设计：从自动化到智能时代的自主化以及自动驾驶车. *应用心理学*, *26*(2), 108–128.

Adams, B. D., Bruyn, L. E., Houde, S., Angelopoulos, P., Iwasa-Madge, K., & McCann, C. (2003). *Trust in automated systems*. Toronto: Ministry of National Defence.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. L. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036–1060.

Armstrong, S. J., Peterson, E. R., & Rayner, S. G. (2012). Understanding and defining cognitive style and learning style: A Delphi study in the context of educational psychology. *Educational Studies, 38*(4), 449–455.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59–64.

Beggiato, M., & Krems, J. F. (2013). The evolution of mental model, trust and acceptance of adaptive cruise control in relation to initial information. *Transportation Research Part F: Traffic Psychology and Behaviour*, *18*, 47–57.

Brown, R. D., & Galster, S. M. (2004, September). Effects of reliable and unreliable automation on subjective measures of mental workload, situation awareness, trust and confidence in a dynamic flight task. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 48, No. 1, pp. 147–151). Sage CA: Los Angeles, CA: SAGE Publications.

Cao, S., Qin, Y. L., & Shen, M. W. (2013). Modeling the effect of driving experience on lane keeping performance using ACT-R cognitive architecture. *Chinese Science Bulletin, 58*(21), 2078–2086.

Carlson, M. S., Drury, J. L., Desai, M., Kwak, H., & Yanco, H. A. (2014, March). Identifying factors that influence trust in automated cars and medical diagnosis systems. In *2014 AAAI Spring Symposium Series*.

Celmer, N., Branaghan, R., & Chiou, E. (2018, September). Trust in branded autonomous vehicles & performance expectations: A theoretical framework. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 62, No. 1, pp. 1761–1765). Sage CA: Los Angeles, CA: SAGE Publications.

Chen, F., Ren, Q., Gao, Z., Wen, Z., & Yang, H. (2021). *Unsettled issues in vehicle autonomy, artificial intelligence, and human-machine interaction*. SAE Technical Paper.

Chen, J. Y. C., Lakhmani, S. G., Stowers, K., Selkowitz, A. R., Wright, J. L., & Barnes, M. (2018). Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical Issues in Ergonomics Science*, *19*(3), 259–282.

Chien, S. Y., Lewis, M., Semnani-Azad, Z., & Sycara, K. (2014, September). An empirical model of cultural factors on trust in automation. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 58, No. 1, pp. 859–863). Sage CA: Los Angeles, CA: SAGE Publications.

Chien, S. Y., Semnani-Azad, Z., Lewis, M., & Sycara, K. (2014, June). Towards the development of an inter-cultural scale to measure trust in automation. In *International conference on cross-cultural design* (pp. 35–46). Springer, Cham.

Chien, S. Y., Sycara, K., Liu, J. S., & Kumru, A. (2016, September). Relation between trust attitudes toward automation, Hofstede's cultural dimensions, and big five personality traits. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 60, No. 1, pp. 841–845). Sage CA: Los Angeles, CA: SAGE Publications.

Choi, J. K., & Ji, Y. G. (2015). Investigating the importance of trust on adopting an autonomous vehicle. *International Journal of Human-Computer Interaction*, *31*(10), 692–702.

Cohen, M. S., Parasuraman, R., Serfaty, D., & Andes, R. C. (1997). *Trust in decision aids: A model and a training strategy*. Arlington, VA: Cognitive Technologies, Inc.

Cramer, H., Evers, V., Kemper, N., & Wielinga, B. (2008, December). Effects of autonomy, traffic conditions and driver personality traits on attitudes and trust towards in-vehicle agents. In *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (Vol. 3, pp. 477–482). IEEE.

de Visser, E. J., Cohen, M., Freedy, A., & Parasuraman, R. (2014, June). A design methodology for trust cue calibration in cognitive agents. In *International conference on virtual, augmented and mixed reality* (pp. 251–262). Springer, Cham.

de Visser, E. J., Pak, R., & Shaw, T. H. (2018). From 'automation' to 'autonomy': The importance of trust repair in human–machine interaction. *Ergonomics*, *61*(10), 1409–1427.

de Vries, P., Midden, C., & Bouwhuis, D. (2003). The effects of errors on system trust, self-confidence, and the allocation of control in route planning. *International Journal of Human Computer Studies, 58*(6), 719–735.

Donmez, B., Boyle, L. N., Lee, J. D., & McGehee, D. V. (2006). Drivers' attitudes toward imperfect distraction mitigation strategies. *Transportation Research Part F: Traffic Psychology and Behaviour, 9*(6), 387–398.

DOT, U. (2018). *Preparing for the future of transportation: Automated vehicles 3.0*. Retrieved December 25, 2020, from https://www.transportation.gov/policy-initiatives/automated-vehicles/av-40

Du, N., Haspiel, J., Zhang, Q. N., Tilbury, D., Pradhan, A., Yang, J., & Robert, L. (2019). Look who's talking now: Implications of AV's explanations on driver's trust, AV

preference, anxiety and mental workload. *Transportation Research Part C: Emerging Technologies, 104*, 428–442.

Ekman, F., Johansson, M., Bligård, L. O., Karlsson, M., & Strömberg, H. (2019). Exploring automated vehicle driving styles as a source of trust information. *Transportation Research Part F: Traffic Psychology and Behaviour, 65*, 268–279.

Ekman, F., Johansson, M., & Sochor, J. (2016, October). To See or Not to See: The Effect of Object Recognition on Users' Trust in" Automated Vehicles". In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction* (pp. 1–4), Gothenburg, Sweden.

Ekman, F., Johansson, M., & Sochor, J. (2018). Creating appropriate trust in automated vehicle systems: A framework for HMI design. *IEEE Transactions on Human-Machine Systems*, *48*(1), 95–101.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 37*(1), 32–64.

Endsley, M. R. (2016). *Designing for situation awareness: An approach to user-centered design* (2ed.). CRC press.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864–886.

European Road Transport Research Advisory Council (ERTRAC) working group. (2019). *Connected Automated Driving Roadmap*. ERTRAC.

Fahrenfort, J. J., van Driel, J., van Gaal, S., & Olivers, C. N. L. (2018). From ERPs to MVPA Using the Amsterdam Decoding and Modeling Toolbox (ADAM). *Frontiers in Neuroscience, 12.*

Forster, Y., Naujoks, F., & Neukum, A. (2017, June). Increasing anthropomorphism and trust in automated driving functions by adding speech output. In *2017 IEEE Intelligent Vehicles Symposium* (pp. 365–372). IEEE.

French, B., Duenser, A., Heathcote, A. (2018). *Trust in Automation – A Literature Review*. CSIRO, Australia.

Gao, J., & Lee, J. D. (2006). Extending the decision field theory to model operators' reliance on automation in supervisory control situations. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 36*(5), 943–959.

Gold, C., Körber, M., Hohenberger, C., Lechner, D., & Bengler, K. (2015). Trust in automation–before and after the experience of take-over scenarios in a highly automated vehicle. *Procedia Manufacturing*, *3*, 3025–3032.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 53*(5), 517–527.

Hancock, P. A., Nourbakhsh, I., & Stewart, J. (2019). On the future of transportation in an era of automated and autonomous vehicles. *Proceedings of the National Academy of Sciences*, *116*(16), 7684–7691.

Hartwich, F., Beggiato, M., & Krems, J. F. (2018). Driving comfort, enjoyment and acceptance of automated driving–effects of drivers' age and driving style familiarity. *Ergonomics, 61*(8), 1017–1032.

Häuslschmid, R., von Bülow, M., Pfleging, B., & Butz, A. (2017, March). Supporting trust in autonomous driving. In *Proceedings of the 22nd international conference on intelligent user interfaces* (pp. 319–329). ACM.

Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding neural representational spaces using multivariate pattern analysis. *Annual Review of Neuroscience, 37*(1), 435–456.

Helldin, T., Falkman, G., Riveiro, M., & Davidsson, S. (2013, October). Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving. In *Proceedings of the 5th international conference on automotive user interfaces and interactive vehicular applications* (pp. 210–217). ACM.

Hergeth, S., Lorenz, L., & Krems, J. F. (2017). Prior familiarization with takeover requests affects drivers' takeover performance and automation trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *59*(3), 457–470.

Hergeth, S., Lorenz, L., Krems, J. F., & Toenert, L. (2015). Effects of take-over requests and cultural background on automation trust in highly automated driving. In *Proceedings of the eighth international driving symposium on human factors in driver assessment, training and vehicle design* (pp. 331–337). University of Iowa.

Hergeth, S., Lorenz, L., Vilimek, R., & Krems, J. F. (2016). Keep your scanners peeled: Gaze behavior as a measure of automation trust during highly automated driving. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *58*(3), 509–519.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(3), 407–434.

Jian, J. Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71.

Jonsson, I. M., Harris, H., & Nass, C. (2008, April). How accurate must an in-car information system be?: Consequences of accurate and inaccurate information in

cars. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1665–1674). ACM.

Kelleher, K. (2018). *Man arrested for drunk driving after officers found him asleep in Tesla running in autopilot mode*. Retrieved December 25, 2020, from http://fortune.com/2018/11/30/man-arrested-drunk-driving-asleep-tesla-autopilot-mode/

Khastgir, S., Birrell, S., Dhadyalla, G., & Jennings, P. (2017). Calibrating trust to increase the use of automated systems in a vehicle. In *Advances in human aspects of transportation* (pp. 535–546). Springer, Cham.

Kohn, S. C., Quinn, D., Pak, R., de Visser, E. J., & Shaw, T. H. (2018, September). Trust Repair Strategies with Self-Driving Vehicles: An Exploratory Study. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 62, No. 1, pp. 1108–1112). Sage CA: Los Angeles, CA: SAGE Publications.

Koo, J., Kwac, J., Ju, W., Steinert, M., Leifer, L., & Nass, C. (2015). Why did my car just do that? Explaining semi-autonomous driving actions to improve driver understanding, trust, and performance. *International Journal on Interactive Design and Manufacturing*, *9*(4), 269–275.

Körber, M., Baseler, E., & Bengler, K. (2018). Introduction matters: Manipulating trust in automation and reliance in automated driving. *Applied Ergonomics*, *66*, 18–31.

Kraus, J., Scholz, D., Stiegemeier, D., & Baumann, M. (2020). The more you know: Trust dynamics and calibration in highly automated driving and the effects of take-overs, system malfunction, and system transparency. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 62(5), 718–736.

Kunze, A., Summerskill, S. J., Marshall, R., & Filtness, A. J. (2019). Automation transparency: Implications of uncertainty communication for human-automation interaction and interfaces. *Ergonomics, 62*(3), 345–360.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors: The Journal of the Human Factors and Ergonomics Society*, *46*(1), 50–80.

Lee, J. G., Kim, K. J., Lee, S., & Shin, D. H. (2015). Can autonomous vehicles be safe and trustworthy? Effects of appearance and autonomy of unmanned driving systems. *International Journal of Human-Computer Interaction*, *31*(10), 682–691.

Li, W. M., Yao, N. L., Shi, Y. W., Nie, W. R., Zhang, Y. H., Li, X. R., ... Gao, Z. F. (2020). Personality Openness Predicts Driver Trust in Automated Driving. *Automotive Innovation*, *3*(1), 3–13.

Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human performance consequences of automated decision aids: The impact of degree of automation and system experience. *Journal of Cognitive Engineering and Decision Making*, *6*(1), 57–87.

Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(2), 194–210.

Meyers, E. (2013). The neural decoding toolbox. *Frontiers in Neuroinformatics*, *7*, 8.

Mirnig, A. G., Wintersberger, P., Sutter, C., & Ziegler, J. (2016, October). A framework for analyzing and calibrating trust in automated vehicles. In *Adjunct proceedings of the 8th international conference on automotive user interfaces and interactive vehicular applications* (pp. 33–38). ACM.

Mishler, S. (2019). *Whose drive is it anyway? Using multiple sequential drives to establish patterns of learned trust, error cost, and non-active trust repair while considering daytime and nighttime differences as a proxy for difficulty* (Unpublished Master's thesis). Old Dominion University, Virginia.

Molnar, L. J. (2017). *Age-related differences in driver behavior associated with automated vehicles and the transfer of control between automated and manual control: A simulator evaluation*. University of Michigan, Ann Arbor, Transportation Research Institute.

Molnar, L. J., Ryan, L. H., Pradhan, A. K., Eby, D. W., Louis, R. M. S., & Zakrajsek, J. S. (2018). Understanding trust and acceptance of automated vehicles: An exploratory simulator study of transfer of control between automated and manual driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, *58*, 319–328.

Morris, D. M., Erno, J. M., & Pilcher, J. J. (2017, September). Electrodermal response and automation trust during simulated self-driving car use. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 61, No. 1, pp. 1759–1762). Sage CA: Los Angeles, CA: SAGE Publications.

National Transportation Safety Board. (2018, March). *Preliminary report: Crash and post-crash fire of electric-powered passenger vehicle*. Retrieved December 25, 2020, from https://www.ntsb.gov/investigations/AccidentReports/Pages/HWY18FH011-preliminary.aspx

Niu, D. F., Terken, J., & Eggen, B. (2018). Anthropomorphizing information to enhance trust in autonomous vehicles. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *28*(6), 352–359.

Noah, B. E., & Walker, B. N. (2017, March). Trust Calibration through Reliability Displays in Automated Vehicles. In *Proceedings of the Companion of the 2017*

ACM/IEEE International Conference on Human-Robot Interaction (pp. 361–362). ACM.

Noah, B. E., Wintersberger, P., Mirnig, A. G., Thakkar, S., Yan, F., Gable, T. M., ... McCall, R. (2017, September). First workshop on trust in the age of automated driving. In *Proceedings of the 9th international conference on automotive user interfaces and interactive vehicular applications adjunct* (pp. 15–21). ACM.

Payre, W., Cestac, J., & Delhomme, P. (2016). Fully automated driving: Impact of trust and practice on manual control recovery. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *58*(2), 229–241.

Petersen, L., Tilbury, D., Robert, L., & Yang, X. J. (2017, August). Effects of augmented situational awareness on driver trust in semi-autonomous vehicle operation. In *2017 NDIA ground vehicle systems engineering and technology symposium.* Novi, MI.

Petersen, L., Zhao, H. J., Tilbury, D. M., Yang, X. J. & Robert, L. P. (2018, August). The influence of risk on driver's trust in autonomous driving system. In *Proceedings of the 2018 ground vehicle systems engineering and technology symposium*, Novi, MI.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., ... Wellman, M. (2019). Machine behaviour. *Nature, 568*(7753), 477–486.

Riding, R., & Cheema, I. (1991). Cognitive styles: An overview and integration. *Educational Psychology, 11*(3–4), 193–215.

SAE China. (2020). *Technology roadmap for energy saving and new energy vehicles 2.0.* China Machine Press.

SAE On-Road Automated Vehicle Standards Committee. (2018). *Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles.* SAE International.

Saffarian, M., de Winter, J. C., & Happee, R. (2012, September). Automated driving: Human-factors issues and design solutions. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 56, No. 1, pp. 2296–2300). Sage CA: Los Angeles, CA: Sage Publications.

Schaefer, K. E., Billings, D. R., Szalma, J. L., Adams, J. K., Sanders, T. L., Chen, J. Y., & Hancock, P. A. (2014). *A meta-analysis of factors influencing the development of trust in automation: Implications for human-robot interaction.* Fort Belvoir, VA: Defense Technical Information Center.

Soh, H., Reid, L. N., & King, K. W. (2009). Measuring trust in advertising. *Journal of Advertising*, *38*(2), 83–104.

Szalma, J. L., & Taylor, G. S. (2011). Individual differences in response to automation: The five factor model of personality. *Journal of Experimental Psychology: Applied*, *17*(2), 71–96.

Toffetti, A., Wilschut, E. S., Martens, M. H., Schieben, A., Rambaldini, A., Merat, N., & Flemisch, F. (2009). CityMobil: Human factor issues regarding highly automated vehicles on eLane. *Transportation Research Record: Journal of the Transportation Research Board, 2110*(1), 1–8.

Verberne, F. M. F., Ham, J., & Midden, C. J. H. (2012). Trust in smart systems: Sharing driving goals and giving information to increase trustworthiness and acceptability of smart systems in cars. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *54*(5), 799–810.

Verberne, F. M. F., Ham, J., & Midden, C. J. H. (2015). Trusting a virtual driver that looks, acts, and thinks like you. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(5), 895–909.

Walker, F., Verwey, W. V., & Martens, M. (2018, June). Gaze behaviour as a measure of trust in automated vehicles. In *Proceedings of the 6th humanist conference*, Hague, Netherlands.

Walker, G. H., & Stanton, N. A. (2017). *Human factors in automotive engineering and technology*. London: CRC Press.

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113–117.

Wintersberger, P., Noah, B. E., Kraus, J., McCall, R., Mirnig, A. G., Kunze, A., ... Walker, B. N. (2018, September). Second Workshop on Trust in the Age of Automated Driving. In *Adjunct proceedings of the 10th international conference on automotive user interfaces and interactive vehicular applications* (pp. 56–64). ACM.

Zhang, Q. N., Robert, L., Du, N., & Yang, X. J. (2018). *Trust in AVs: The impact of expectations and individual differences*. Presented at the Conference on Autonomous Vehicles in Society: Building a Research Agenda, Ann Arbor, MI.

Zihsler, J., Hock, P., Walch, M., Dzuba, K., Schwager, D., Szauer, P., & Rukzio, E. (2016, October). Carvatar: Increasing trust in highly-automated driving through social cues. In *Adjunct proceedings of the 8th international conference on automotive user interfaces and interactive vehicular applications* (pp. 9–14). ACM.

Trust in automated vehicles

GAO Zaifeng[1], LI Wenmin[1], LIANG Jiawen[1], PAN Hanxi[1], XU Wei[2], SHEN Mowei[1]

([1] *Department of Psychology and Behavioral Sciences, Zhejiang University*;

[2] *Center for Psychological Sciences at Zhejiang University, Hangzhou* 310007, *China*)

Abstract: Automated driving (AD) is one of the key directions in the intelligent vehicles field. Before full automated driving, we are at the stage of human-machine cooperative driving: Drivers share the driving control with the automated vehicles. Trust in automated vehicles plays a pivotal role in traffic safety and the efficiency of human-machine collaboration. It is vital for drivers to keep an appropriate trust level to avoid accidents. We proposed a dynamic trust framework to elaborate the development of trust and the underlying factors affecting trust. The dynamic trust framework divides the development of trust into four stages: dispositional, initial, ongoing, and post-task trust. Based on the operator characteristics (human), system characteristics (automated driving system), and situation characteristics (environment), the framework identifies potential key factors at each stage and the relation between them. According to the framework, trust calibration can be improved from three approaches: trust monitoring, driver training, and optimizing HMI design. Future research should pay attention to the following four perspectives: the influence of driver and HMI characteristics on trust, the real-time measurement and functional specificity of trust, the mutual trust mechanism between drivers and AD systems, and ways in improving the external validity of trust studies.

Key words: trust, automated driving, dynamic trust framework, trust calibration, HMI design